一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-046-09 (11/09)

Using utterance timing to generate gaze pattern*

Jani Even, Carlos Toshinori Ishi, Hiroshi Ishiguro
Hiroshi Ishiguro Laboratories, Advanced Telecommunications Research Institute International, Japan.
even@atr.jp *

Abstract

This paper presents a method for generating the gaze pattern of a robot while it is talking. The goal is to prevent the robot's conversational partner from interrupting the robot at inappropriate moments. The proposed approach has two steps: First, the robot's utterance are split into meaningful parts. Then, for each of these parts, the robot performs or avoids eyes contact with the partner. The generated gaze pattern indicates the conversational partner that the robot has finished talking or not. To measure the efficiency of the approach, we propose to use speech overlap during conversations and average response time. Preliminary results showed that setting a gaze pattern for a robot with a very human-like appearance is not straight forward as we did not find satisfying parameters.

1 INTRODUCTION

During social interaction, the gaze has important regulatory functions [1, 2]. Early work [1, 3] tried to search a systematic relation between gaze and turn-taking. More recent work [4] underlines the collaborative nature of the gaze in turn-taking. The authors in [4] show that "the timing of the listener response is collaborative process, accomplished by joint action". Consequently, a robot holding a conversation with a human should participate in this "collaborative process" in order to have a smooth interaction.

This paper presents an approach to robot gaze control during conversation that take into account this collaborative process. The goal is to make the conversation flow smoother by providing the human with the expected gaze signals that occur during turn-taking.

An expected outcome is to reduce the risk that the listener interrupts the robot. Without signaling, it is quite frequent that the human interrupts the robot because she or he did not understand that the robot intended to continue speaking.

Figure 1: Close-up of Erica.

Using gaze to signal the turn taking is also expected to avoid undesired pauses caused by the listener not taking it's turn fast enough.

In addition to provide the adequate signaling, the gaze pattern should not be unnatural. The implementation of a human-like solution to the problem of unwanted interruption is all the more important as we work with an android robot developed to look very similar to human. This appearance similarity amplifies the sensitivity to unnatural behaviors.

For this reason, the proposed gaze pattern should also display the same habit as human to avert gaze during utterance formulation [5]. Humans tend to avert their gaze during formulation to reduce the cognitive load. This a natural phenomenon we are used to witness during a conversation. Thus, in addition to the turn taking signal, the proposed gaze pattern generation also tries to reproduce the aversion that occurs during formulation.

2 RELATED WORK

The use of gaze by social robot during interaction have been investigated by several authors with different per-

*Research supported by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

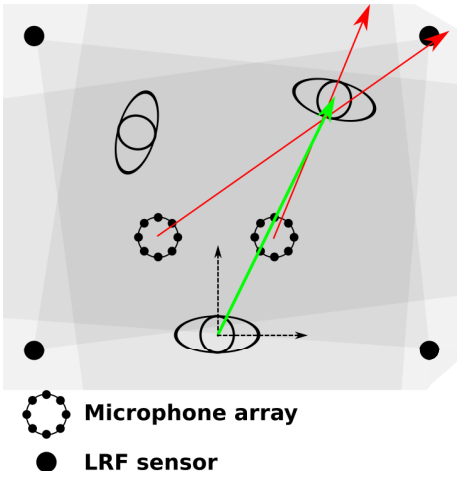

Figure 2: Example of possible sensor network configuration.

spectives. In particular, in [6], the authors recorded human-human conversations in order to estimate statistics from the gaze patterns. Then, these statistics were used to implement the gaze control of a NAO robot. The NOA robot was judged by participants to be more thoughtful and it was able to manage the conversation floor. In this paper, we would like to achieve similar results using a robot that is more human-like than NAO.

3 ROBOT GAZE CONTROL

The proposed gaze pattern generation is designed for Erica [7], a robot that was designed to have a realistic human like appearance, see Fig.1.

The components of Erica that are involved in the gaze control are:

- a sensor network,
- a kinematic model,
- and a closed-loop controller.

The sensor network main role is to track human [8, 9, 10] and determine who is talking [11, 12]. For this purpose a human tracking system is combined with a sound localization system. Figure 2 shows one example of configuration with four laser range finders (LRFs) for tracking humans and two microphone arrays for performing sound localization. During the experiments, the human tracker system was not using LRFs but RGB-D cameras attached to the ceiling of the room [13]. Using the sound localization (the red arrows in Fig.2) it is possible to determine who is talking.

Figure 3 shows the joints involved in the gaze control. The kinematic chain controlling the eyes direction has 7 degrees of freedom (DOF):

- yaw and pitch for the eyes,
- yaw pitch and roll for the neck,
- yaw and pitch for the waist.

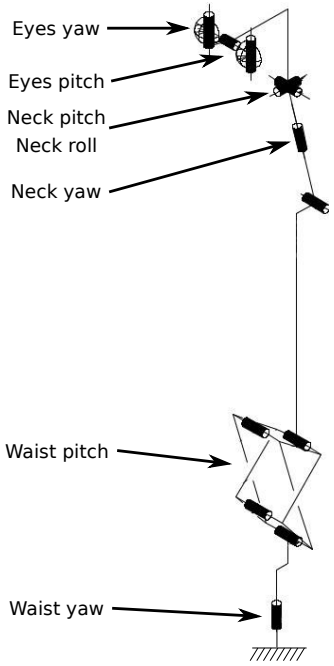

Figure 3: Kinematic chain for the gaze.

However, the current implementation does not use the neck roll.

Pneumatic actuators are used to move the joints. These actuators are controlled by on board PID controllers. The commands are sent to the robot at a frequency of 20 Hz. The robot provides a feedback measured by potentiometers also at the frequency of 20 Hz. The on board PID are tuned to favor smoother movements which results in a lesser control accuracy. Consequently, it is necessary to rely on the feedback to get the achieved positioning.

Using the specifications of Erica, a computer model of the kinematic chain was implemented. The posture of the model is updated when the feedback from the actuators is received. Namely, the model provides an estimate of the current posture of Erica.

The kinematic model provides the current gaze direction of Erica's eyes. The goal of the gaze control is to send command to move the joints of Erica in order to align Erica's gaze direction to the desired gaze direction. Only the eyes are controlled in a closed loop because the accuracy on the eye movement is greater than on the waist and neck.

Erica is able to track a moving person walking in front of her using the gaze control. This is illustrated in Fig.4. The top of Fig.4 shows the yaw of the focus direction (solid line) and the yaw of the gaze direction given by the kinematic model (dashed line). The three other graphs are showing the command values (solid lines) and the potentiometer values (dashed lines) for the control of the waist, neck and eyes yaw. We can note a slight delay, which is expected, and some overshoots. However, the graph for the neck control shows some large errors and the one for the waist some small errors. Then, we can see on the graph for the eyes that the command is different

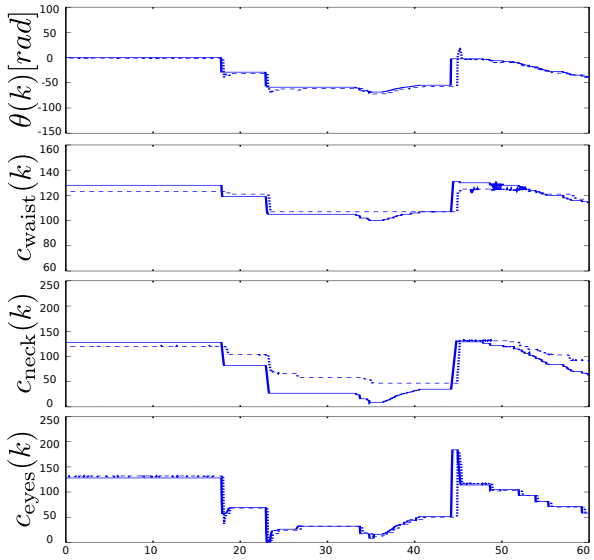

Figure 4: Close-up of the axis command (dashed) and potentiometer feedback (solid) for the yaw.

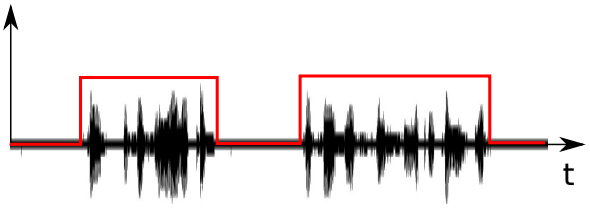

Figure 5: Robot utterance and voice activity (red).

and it compensated for the error as expected.

4 ROBOT GAZE PATTERN

Previous section showed that Erica is able to look relatively precisely to a given direction. In this section, we will discuss the gaze pattern during interaction. In particular, the focus is on the gaze pattern when the robot is talking.

4.1 Gaze pattern timing

Figure 5 shows a typical utterance of the robot. It is possible to have a precise voice activity detection for the robot speech from the text to speech (TTS) module.

The goal is to produce a gaze pattern that presents the adequate cognition and turn taking cues. Such a gaze pattern is illustrated in Fig. 6. The gaze pattern is a succession of gaze aversion and eye contact that have timed in a specific manner.

In particular, if a single utterance is considered, as in Fig. 7, it is split in different phases:

- The cognition phase, denoted by C, starts before the utterance and overlaps the beginning of the utterance. This phase corresponds to the duration during which gaze aversion is expected as the talking would be formulating her or his utterance.
- The final phase, denoted by F, starts before the end of the utterance and persists after the utterance is over. During this phase, the turn taking signal is sent to the conversation partner. If the speaker wants to give the turn, eye contact is sought. But, on the contrary, if the speaker intends to continue talking the gaze is averted.
- The body phase, denoted by B, is the duration between the end of the cognition phase and the beginning of the final phase. During this phase, a succession of short gaze aversions and eye contacts occurs.

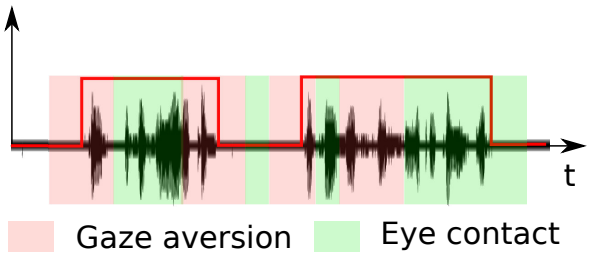

Figure 6: Gaze pattern for two consecutive utterances.

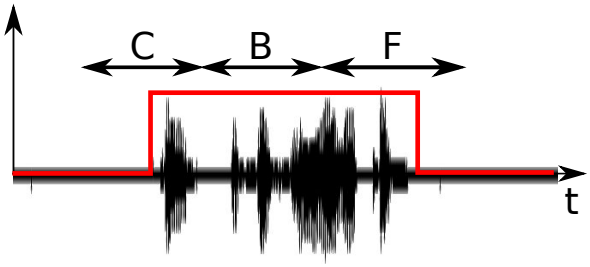

Figure 7: The three utterance phases.

In order to produce the gaze pattern, it is necessary to anticipate the start of the utterance to be able to perform the aversion of the cognitive phase and the end of the utterance to be able to start signaling the turn taking in advance.

The sequence of actions that results in the robot speaking is indicated on the time line of by Fig. 8:

- a speak request is sent at time S to the TTS module,
- the synthesized speech is ready at time T,
- the lip synchronization commands are sent to the robot at time C,
- The lip movement and the sound production start at time V.

The delay d between speech request and actual speech production is due to the necessity to synchronize the speech sound with the lip motion. Because of the pneumatic actuation, the shortest possible delay is 500 ms. The expected end of the utterance is known in advance as the duration D of the utterance is deduced from the synthesized speech. Thus, it is possible to access in advance the voice activity of the robot and all the timing information necessary to create the gaze pattern are readily

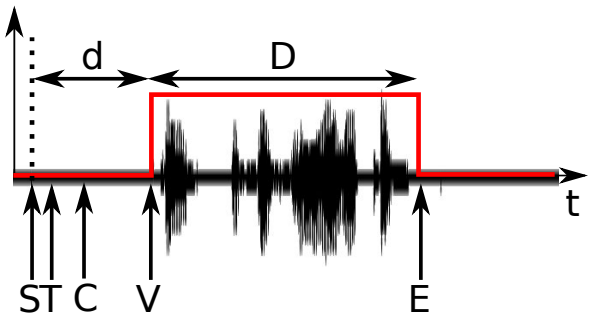

Figure 8: Sequence of events during an utterance production.

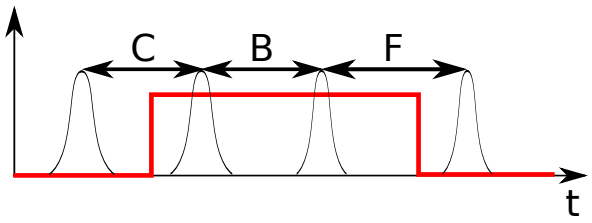

Figure 9: Sampling phase boundaries from Gaussian distributions.

available. Note that to generate a cognition phase longer than 500 ms, it is necessary to add a pre-delay.

In order to have a more natural gaze pattern, it is important to introduce some variability in the different phase durations. In the current implementation, the start time and end time of the different phases are sampled from Gaussian distributions. Figure 9 shows that the phases are determined by four distributions.

In the body phase, the gaze pattern is composed of short aversions and eye contacts. This is done by sampling the aversion duration and the duration between successive aversions from two Gaussian distributions.

Note that when the robot is listening, the gaze pattern is also a succession of aversions and eye contacts. However, the duration of the aversions and the intervals between them are shorter. These patterns are generated this way as humans also tends to do shorter and less frequent aversion when listening than when talking [1].

4.2 Gaze pattern direction

The eye contact is performed by having the robot look at the human that is detected by the sensor network. The sensor network gives the position of the person in the room and the height of that person (the top of the head). In the current implementation, the gaze controller set the robot to look at a fixed offset of 0.15 meters from the top of the head.

During gaze aversion, an offset is added to the gaze controller that results in the eyes of the robot being averted from the person. The gaze aversion offset is characterized by two angles θ in the horizontal plane and ϕ in the vertical plane.

To introduce randomness in the gaze aversion, first the "general direction" is selected among "up", "down",

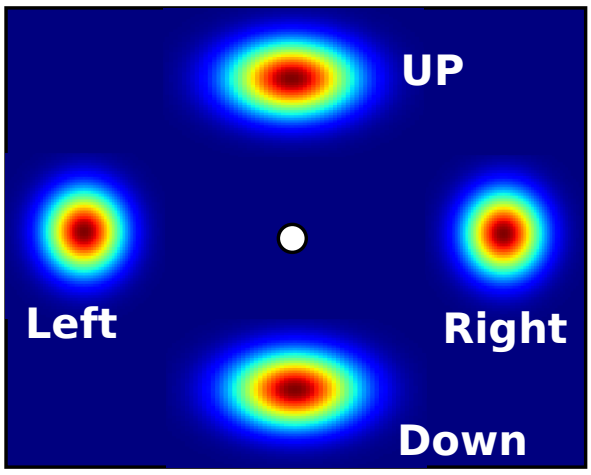

Figure 10: Sampling gaze directions from Gaussian distributions in the four "general directions".

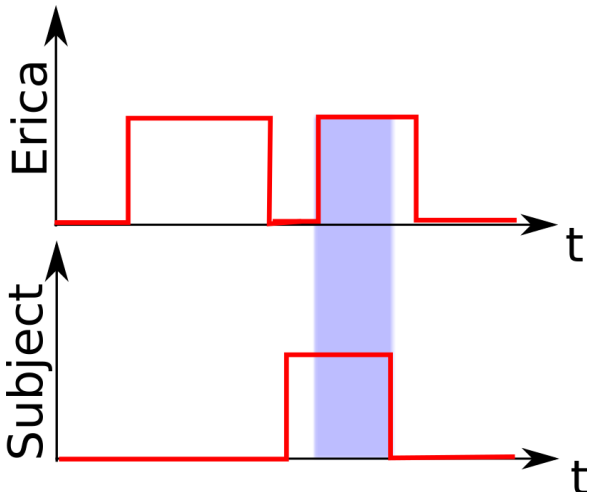

Figure 11: Speech overlap caused by missing the signal indicating a short pause.

"left" or "right". Then, the magnitude of the aversion is sampled from Gaussian distributions, see Fig. 10.

5 EXPERIMENTAL SYSTEM

To asses the effect of the gaze pattern on the conversation, it is necessary to define a measure of performance. For this purpose, a conversation monitoring system is used. Using the microphone arrays of the sensor network, the speech activity of the subject conversing with Erica is logged. The utterance timing and the gaze pattern information are also recorded at the same time.

Figure 11 shows an overlap situation. The subject did not understand that Erica was just making a short pause and took the talking turn. Even if the subject voice is detected it is usually too late to avoid a slight overlap as the robot commands are sent in advance. This is a typical case where the robot should use gaze aversion to signal the turn is not over. Thus, a measure of performance is the duration of the overlap (the blue region in Fig. 11).

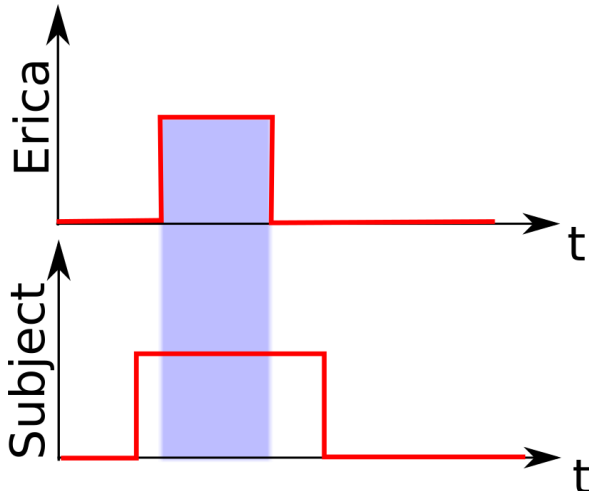

Figure 12: Speech overlap caused by missing the signal indicating the robot is about to talk.

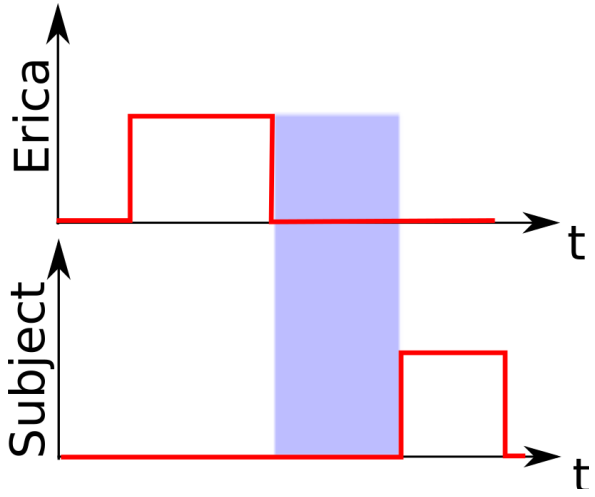

Figure 13: Example of delayed answer.

Another overlap situation is the one when the subject could not anticipate that the robot was about to talk. Then, the subject starts talking just before the robot resulting in an overlap. In this case, illustrated in Fig. 12, not signaling the cognition phase by a gaze aversion resulted in the overlap (in blue).

In addition to speech overlaps, the conversation is not smooth when the subject takes much time to answer to the robot because the robot did not signal properly the end of turn. Thus, measuring the response time of the subject, the blue duration in Fig. 13 is another good indication of performance.

For a given conversation between a subject and the robot, it is possible to calculate the total speech overlap and the average response time. These two values are used as the measure of performance for the generated gaze pattern. Namely, between two candidate gaze patterns the one that results in less speech overlap and a smaller average response time is considered better.

In order to compare different gaze patterns for different subjects, the random generation of the phase timing are only computed one time and stored. Then, it is possible to compare the same realization of the gaze pattern by replaying the stored version to each subjects.

Preliminary experiments have showed that creating the gaze pattern using Gaussian distribution is not straight forward. The first attempts used the means and standard deviations that were estimated from human-human conversation in [6]. However, the generated gaze patterns were not satisfying. In particular, the cognition and final phase tended to be too long for the robot utterances. The reason is maybe that the subjects in [6] were familiar with each others and had conversations composed of rather long utterances. In comparison, the robot-subject utterances tend to be shorter. Another possible explanation is cultural difference as the results reported in [6] are for English whereas we conducted our experiments in Japanese. It is also possible that we are more sensitive to the gaze pattern mismatch as Erica is more human-like than NAO.

6 CONCLUSIONS

In this paper, we motivated the need for a gaze pattern generation that helps the robot to have a smoother conversation. We presented the architecture of the system and explained the concept of the gaze pattern generation. However, the implementation and testing of the system was not done yet as preliminary results showed that finding a reasonable set of parameters for our robot is not as simple as expected. Thus, the focus now is on adapting the gaze pattern statistics to our specific robot in order to proceed with the evaluation. A possibility is to generate the statistics by taking into account the duration of the utterance when creating the gaze pattern.

References

[1] Kendon A., "Some functions of gaze-direction in social interaction," *Acta Psychol.*, vol. 26, no. 1, pp. 22–63, 1967.

[2] M. Argyle and M. Cook, *Gaze and mutual gaze*, Cambridge University Press, 1976.

[3] S. D. Jr. Duncan, "Some signals and rules for taking speaking turns in conversation," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[4] J. B. Bavelas, L. Coates, and T. Johnson, "Listener responses as a collaborative process: The role of gaze," *Journal of Communication*, vol. 52, no. 3, pp. 566–580, 2002.

[5] A.M. Glenberg, J.L. Schroeder, and D.A. Robertson, "Averting the gaze disengages the environment and facilitates remembering," *Memory and Cognition*, vol. 26, pp. 651–658, 1998.

[6] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu, "Conversational gaze aversion for humanlike robots," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, 2014, pp. 25–32.

[7] Hiroshi Ishiguro et al., "Erato ishiguro symbiotic human-robot interaction project," `http://www.jst.go.jp/erato/ishiguro/en/index.html`, 2015.

[8] Jae Hoon Lee, T Tsubouchi, K Yamamoto, and S Egawa, "People tracking using a robot in motion with laser range finder," 2006, pp. 2936–2942, Ieee.

[9] D.F. Glas et al., "Laser tracking of human body motion using adaptive shape modeling," *Proceedings of 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 602–608, 2007.

[10] L. Spinello and K. O. Arras, "People detection in rgb-d data.," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[11] C.T. Ishi et al., "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," *Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2027–2032, 2009.

[12] C.T. Ishi, J. Even, and N. Hagita, "Using multiple microphone arrays and reflections for 3d localization of sound sources," *Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3937–3942, 2013.

[13] D. Brscic, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *Human-Machine Systems, IEEE Transactions on*, vol. 43, no. 6, pp. 522–534, 2013.